DOCUMENT RESUME

ED 119 634 IR 003 095

AUTHOR Digger, J.
TITLE The Provision of Subject Data on EUDISED Records.
INSTITUTION Council for Cultural Cooperation, Strasbourg (France). Ad Hoc Committee for Educational Documentation and Information.
REPORT NO DECS-DOC-75-33
PUB DATE 5 Dec 75
NOTE 31p.; Not available in hard copy due to marginal reproducibility of original document

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Classification; Data Bases; *Indexing; *Information Retrieval; *Information Services; Information Storage; Information Systems; International Organizations; Relevance (Information Retrieval); Search Strategies; *Subject Index Terms; Thesauri
IDENTIFIERS *EUDISED

ABSTRACT

A study was conducted to determine the most effective way to organize the data that are being accumulated by the European Documentation and Information System for Education (EUDISED). The study considered types of information services and search facilities which might be used to access the EUDISED data base. Indexing systems were examined and their effectiveness compared for various kinds of search tasks. A possible method for providing a multilingual browsing facility was suggested, and a tentative strategy was outlined for providing EUDISED records with subject data. (EMH)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED119634

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Strasbourg, 5 December 1975

DECS/Doc (75) 33
Engl. only

I R

COUNCIL FOR CULTURAL CO-OPERATION

Committee for Educational
Documentation and Information

THE PROVISION OF SUBJECT DATA ON EUDISED RECORDS

A paper prepared by Mr. J. Digger, Subject Systems Office,

British Library, Bibliographical Services Division

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

42.364
04.6

ER 003 095

THE PROVISION OF SUBJECT DATA ON EUDISED RECORDS

1. Terminology

1.1 The word 'data' is now frequently treated as a singular English noun rather than as a Latin plural (see, for instance, many of the contributions to 'Encyclopaedia of linguistics, information and control' [1]). This usage will be followed throughout this study.

1.2 Mills[2] has drawn a convenient distinction between two broad classes of subject data :

a) data which gives a direct indication of a document's subject content, e.g. classmarks, controlled terms selected from a thesaurus, 'free language' terms extracted from a document's text

b) data which only indicates subject content in an oblique or indirect manner, e.g. the bibliographic citations appended to a document, which are sometimes taken - at least, for the purposes of citation indexing - as being subject-indicative. Only the types of subject data falling within category (a) above will be considered here.

1.3 The expression 'indexing system' will here be used in a general sense, to signify any system (of whatever kind - classification, thesaurus, etc.) for providing documents with subject data.

2. The Scope of the Present Study

2.1 The overall aim of this study is to make certain tentative suggestions as to the types of subject data which might be provided on EUDISED records, so as to enable these records to be used as the basis for a wide range of effective information services.

2.2 Strictly, it is not possible to discuss the provision of subject data on machine-readable records without some thought as to how this data will be created and used. This raises such matters as indexing

policies (e.g. as to exhaustivity of indexing), data processing capabilities, and search techniques. Since these topics cannot be dealt with in detail here, they will only be mentioned where it is felt that they have a direct bearing on the decision as to the type of subject data best fitted to a particular purpose.

3. The Conceptual Approach

3.1 This study will first set out to identify the types of information service and search facility which might be offered as a modern information system utilising the EUDISED data base. A distinction will be drawn between 'retrieval searches' and 'browsing searches'.

3.2 Various types of indexing system will be defined, and compared in respect of their performance potential for retrieval search. The comparison will be made on the basis of six performance criteria.

3.3 A possible method for providing a multilingual browsing facility in bibliographic tools will be suggested. This leads to a general consideration of the problems of providing multilingual access in information systems. Finally, a tentative strategy is outlined for providing EUDISED records with subject data.

4. Information Services Based on EUDISED Records

4.1 The first requirement is to identify the types of information service which might make use of subject data on EUDISED records. According to Thompson[3] an information system can be regarded as having three main functions : current awareness, retrospective search and the compilation of literature surveys. The last of these functions will be disregarded here, since it does not directly involve subject data. The main types of current awareness and retrospective search services one would expect to find in a modern computer-based information system are as follows :

a) An On-line Search Service

In order to keep computing costs and telecommunication charges within reasonable bounds, a limit is normally set on the number of citations a user may retrieve on-line. Often, only part of the database is accessible on-line - the older portion of the file being available solely for off-line search.

(A typical example of this approach is to be found in MEDLARS (MEDical Literature Analysis and Retrieval System) which now has a data base of more than 2,000,000 records. At the time of writing, only about 600,000 of these records can be accessed through MEDLINE (MEDLARS on-line)[4]. UK MEDLARS restricts the output for an on-line search to a maximum of 25 citations)

b) An Off-line Retrospective Search Service

Off-line searching is applicable when :

i) a user is unable to search on-line - perhaps through lack of access to appropriate telecommunication facilities.

ii) a comprehensive search is required, involving back-files which are not available on-line.

iii) the number of citations to be retrieved exceeds the permitted limit for on-line searches.

c) Selective Dissemination of Information (SDI)

The services so far described primarily cater for 'one-off' searches. SDI provides for repetitive searching in which user profiles are matched at regular intervals, against the most current portion of the database. Profile matching is normally performed as an off-line batch processing operation.

d) Recurrent Bibliographies

These might be of two types :

i) a bibliography covering the total intake to the database. Such a bibliography would typically be issued quite frequently (e.g. monthly), with occasional cumulations (say, six-monthly and annual) (cf.'Current Index to Journals in Education')

ii) bibliographies covering certain defined subsets of the total database, e.g. documents of a particular form (say, audio-visual materials); documents in a particular subject area (say, 'The education of the handicapped') (cf. the special bibliographies produced by some ERIC (Educational Resources Information Center) clearinghouses).

e) Current Awareness Bulletins

Like the recurrent bibliographies, these might be of two types :

i) a bulletin covering all recent additions to the database.

ii) separate bulletins covering various types of recent additions

Current awareness bulletins are normally issued with high frequency (say, weekly or fortnightly).

5. Types of Search

5.1 Searches will here be divided into 'retrieval searches' and 'browsing searches'. The former are conventional literature searches in which the user has at least some notion - albeit a vague one - of the type of document he is seeking. 'Browsing' characterises the behaviour of a user whose search philosophy might be summed up as, 'I don't know what I'm looking for, but I shall know it when I see it'.

6. The Relationship Between Types of Search and Types of Service

6.1 The table below relates type of search to type of information service.

		TYPES OF SEARCH	
TYPES OF SERVICE		Retrieval searches	Browsing
On-line search service		✓	
Off-line search service	Retrospective	✓	
	SDI	✓	
Recurrent bibliographies		✓	✓
Current awareness bulletins			✓

A '✓' indicates that a service makes provision for a particular type of search, a blank indicates that it does not.

6.2 Some explanation is required of the grounds for assuming in the table that most information services only cater for a particular type of search.

a) On-line Searching

It would be uneconomical - in terms of computer time and telecommunication charges - to allow protracted browsing searches to be conducted on-line.

b) Off-line Searching (retrospective, and SDI)

Browsing is obviously impossible when searching off-line, since neither the index nor the citations file is visible to the searcher during the search process.

c) Current Awareness Bulletins

On the assumption that :

i) current awareness bulletins are published sufficiently frequently to ensure that each issue contains only a relatively small number of citations,

ii) these citations are arranged in subject groupings,

a user should be able to identify possibly relevant citations by browsing quickly through the appropriate section(s) of a bulletin. The implication is that current awareness bulletins need not provide for retrieval searches, and so can dispense with subject indexes. This approach seems particularly justified if bulletins are later to be replaced by recurrent bibliographies with full subject indexes.

6.3 At this point, we shall temporarily set aside any further mention of browsing searches; this topic will be taken up again in Section 10. Sections 7-9 are given over to a consideration of the types of subject data appropriate to retrieval searches.

7. Performance Criteria Applicable to Retrieval Searches

7.1 Information systems are generally judged in terms of :

a) effectiveness, i.e. the degree to which they satisfy users' requirements.

b) efficiency, i.e. the degree to which they satisfy the management requirements that they be economical to establish and operate

7.2 Effectiveness Criteria

7.2.1 A system is normally deemed effective if it performs well in the following respects :

Recall
Precision
Coverage
Currency
Response time
Ease of use
Appropriateness of form of output

This is a somewhat modified and extended version of the well known list of performance criteria suggested by Lancaster[5]. Since a system's 'coverage', 'currency' and 'appropriateness of form of output' are not directly dependent upon the types of subject data it uses these criteria will be ignored here. This leaves 'Recall', 'Precision', 'Response time' and 'Ease of use' for further consideration.

7.2.2 'Recall' and 'Precision' are here used in a general sense as what Snyder[6] has called 'criterion concepts' rather than as the names of particular quantitative measures of retrieval performance. A system's recall is its ability to retrieve relevant citations, its precision is its ability to avoid the retrieval of non-relevant citations. The notions of recall and precision seem applicable to the whole spectrum of retrieval searches, though not to browsing, since this is something of a 'lucky dip' search technique in which neither high recall nor high precision is expected.

7.2.3 'Response time' will here be interpreted as 'search effort', since this is the component of response time which is particularly dependent upon the nature of the subject data a system provides.

7.2.4 'Ease of use' is a subjective factor which varies from user to user, and which is likely to be at least partly reflected in search time, i.e. there is probably a strong correlation between ease of use and speed of use. However, one aspect of 'ease of use' warrants separate consideration : this is the degree to which a system may be used without the need for the user to comply with a variety of special system-imposed protocols and conventions. Watt[7] has called this property of a system its 'habitability'. Although this term is somewhat unfamiliar it will be adopted here in the absence of any obviously better alternative.

7.3 Efficiency Criteria

7.3.1 The following efficiency criteria are particularly relevent to information systems providing retrieval facilities :

Indexing effort
Vocabulary maintenance effort

Efficiency improves as the effort devoted to indexing and vocabulary maintenance decreases.

7.4 Search effort, habitability, indexing effort, and vocabulary maintenance effort are all, in an obvious sense, subsidiary to recall and precision, since unless a system has at least some success in retrieving relevant documents and avoiding non-relevant ones, its performance in other respects is only of academic interest.

8. Types of indexing system

8.1 This section will identify various types of indexing system which might be used to provide subject data for EUDISED records. Systems will here be discussed in terms of :

a) the method of term co-ordination they use : pre-co-ordinate or post-co-ordinate.

b) their type of vocabulary

8.2 Pre- and post- co-ordinate systems

8.2.1 Pre-co-ordinate systems

In a pre-co-ordinate system, terms are co-ordinated (i.e. combined) at the time of indexing, so that each index entry serves as a kind of 'telegraphic' statement of the subject indexed e.g.

Students. Universities. Great Britain
Attitudes to curriculum - Surveys 174

Pre-co-ordinate index entries of this type show the various 'contexts' in which each lead term in the index (i.e. 'Students' in the example above) has occurred. This contextual information helps the user to decide whether or not the documents to which a particular entry refers are likely to be relevant to his enquiry.

8.2.2 Pre-co-ordinate systems which use controlled vocabularies are conventionally further characterised as being either synthetic or enumerative. In synthetic systems, compound subjects are specified by selecting terms from a thesaurus or classification schedule and combining these, according to a preferred citation order, into a pre-co-ordinated 'string'. Depending upon the type of vocabulary employed, this stage will normally be more or less subject-co-extensive. An enumerative system, on the other hand, attempts to supply ready-made subject headings (or class numbers) for all subjects which might form the focus of a user's enquiry. In practice the headings provided tend to be considerably less specific than many of the subjects they are required to convey - so much so that, even for monograph indexing, the adequacy of enumerative systems is now seriously in doubt[8]. For this reason systems of this type will not be discussed further in the context of retrieval searches (though it will later be suggested that they may have some merit as 'browsing schemes'). Their inability to provide satisfactory search keys for the wide range of highly-specific subject likely to occur in EUDISED materials is taken to be self-evident. The expression 'pre-co-ordinate system' will henceforth be used to mean a 'synthetic pre-co-ordinate system',

8.2.3 Post-co-ordinate systems

In a post-co-ordinate index, each term appears in isolation, e.g.

Attitudes	Curriculum	Great Britain
174	174	174
Students	Surveys	Universities
174	174	174

and although it is possible to combine terms at the time of search, e.g.

Students and Attitudes

the searcher cannot tell the various contexts in which any particular combination of terms has occurred.

8.2.4 The use of contextual information, as a means of making relevance judgements in the course of a search, depends upon the searcher's ability to interact with the index. This interaction is only possible if the index is visually displayed - as in an on-line search, or a search of a bibliography. Where the index is not visible to the searcher - as in the case of an off-line search - all searches have of necessity to be conducted in a post-co-ordinate manner. It should be noted that whereas a post-co-ordinate search may be performed in a pre-co-ordinate index, simply by ignoring the contextual information the index provides, post-co-ordinate indexes cannot be treated as though they were pre-co-ordinate. It is not possible, at the time of search, to arrange unordered lists of terms into pre-co-ordinated 'strings' without the fear that the resulting index will contain many entries which inaccurately or ambiguously represent the subject they ought, ideally, to convey.

8.3 Types of vocabulary

8.3.1 Uncontrolled vocabularies

Included in this category are :

a) vocabularies consisting of terms selected (either manually or automatically) from titles, abstracts, full text, or any combination of these.

b) vocabularies consisting of all the words appearing in the titles of documents, or in their titles and abstracts. Retrieval activities based on this kind of vocabulary are normally referred to as 'free-text searches' - the technique employed, for instance, in IBM's ITIRC system[9]. The possibility of searching the full texts of documents, as is done, for instance, in some legal text retrieval systems such as STATUS[10]) will not be considered here. It will be assumed that, for cost reasons alone, it would be totally impracticable to convert the full text of every document in the EUDISED data base to machine-readable form, store this data, and provide a random access search facility on all significant words.

c) vocabularies consisting of terms freely assigned by indexers working without the constraints of any kind of authority list of terms, e.g. the 'free indexing' practised by INSPEC[11]).

8.3.2 Open-ended controlled vocabularies

Systems of this type :

a) provide terms which are co-extensive with the concepts they are intended to convey. Since, in the course of time, new concepts emerge in the literature of a subject field, adherence to the philosophy of maximum specificity in indexing and searching implies the use of an open-ended vocabulary to which new terms can be added as and when new concepts are encountered by the indexer.

b) introduce controls into the vocabulary to avoid the semantic scatter which occurs if the same concept is expressed by two or more different terms. Where two or more terms are equivalent in meaning for retrieval purposes, i.e. they represent the same concept (e.g. 'Employees','Staff','Personnel') one is chosen as the preferred term (say 'Personnel'). The other terms are given the status of 'Forbidden'terms, and 'See'or 'Use' references made from them to the preferred term. This procedure is calculated to promote recall by ensuring that indexers and searchers achieve a'coincidence of

8.3.3 Fixed vocabularies

Vocabularies of this type are controlled but, by comparison with those of the open-ended kind are relatively fixed, i.e. they are only occasionally updated, when new editions of the vocabulary are published. In consequence newly-emergent concepts have sometimes to be expressed by whatever term represents the 'nearest generic head'. This practice simplifies vocabulary maintenance, but introduces a danger of loss of precision in retrieval.

Non-specific controlled vocabularies range from the 'almost specific' to the 'very broad'. Those at the broad end of the spectrum - such as the limited vocabularies at one time in vogue amongst post-co-ordinate feature card systems - will be disregarded here. Vocabularies of this type generally perform with low precision (see, for instance,[12]), and are therefore poorly equipped to provide for retrieval searches in large-scale computer-based information systems.

8.4 The two factors described above - method of co-ordination and type of vocabulary - can be combined to produce a classification of six types of indexing system :

	METHOD OF TERM CO-ORDINATION	
TYPE OF VOCABULARY	Pre-co-ordinate	Post-co-ordinate
Uncontrolled vocabularies	e.g. title-based KWIC indexes	e.g. free text
Open-ended controlled vocabularies	e.g. PRECIS [13]	e.g. Excepta Medica system[14]
Closed vocabularies	e.g. faceted classification schemes	e.g. EUDISED thesaurus[15]

9. The performance potential of various types of indexing system

9.1 In Section 6.1, it was suggested that information systems commonly made provision for retrieval searches in the following situation : on-line searching, off-line retrospective searching, SDI, and the searching of recurrent bibliographies. These four situations can conveniently be reduced to two :

a) post-co-ordinate search situation. These are the situations in which the index is, of necessity, 'hidden' from the user, so that all searching must be performed in a post-co-ordinate manner, ignoring any contextual information provided by pre-co-ordinate index entries.

b) searches of visually-readable indexes, i.e. situation in which either pre- or post-co-ordinate searching is possible; although, as previously noted (Section 8.2.4) pre-co-ordinate searching presupposes the availability of a pre-co-ordinate index.

Off-line searches whether, of the retrospective or SDI variety, are clearly of type (a), while searches of recurrent bibliographies qualify as type (b). On-line searches may be of type (a) or (b), depending upon the search facilities available. At present, most on-line systems provide only for post-co-ordinate searching, though there is no reason in principle why they should not allow the on-line display of pre-co-ordinate index entries.

9.2 The types of indexing system defined in the last section will now be compared in respect of their ability to provide EUDISED records with suitable subject data. The comparison will involve an assessment of the performance potential of the various systems, both in post-co-ordinate search situations, and when searching visually-readable indexes. Performance will be judged in terms of the six criteria previously defined (see Section 7.4).

9.3 Recall

9.3.1 The retrieval tests conducted to date suggest that a system's recall performance is not significantly affected by the nature of the terms in its vocabulary, e.g. their specificity or method of co-ordination. Given adequate indexing and searching decisions any system can attain good recall, provided that its vocabulary incorporates the necessary 'recall devices'. Recall devices - such as the confounding of synonyms, or the display of hierarchical relationships - allow searches to be expanded so that they retrieve more documents. The expansion is accomplished in a systematic manner so that there is a high probability that at least some of the additional documents captured will be relevant to the user's enquiry, and that recall will thereby be improved.

9.3.2 Recall devices may be mandatory, e.g.

Staff See Personnel

or optional, e.g.

Schools
See also
Secondary schools

In principle, all systems may embody the same recall devices, but different systems are forced to adopt these devices in different ways. For instance, in an uncontrolled vocabulary recall devices are always optional (since to make any of them mandatory would be to introduce an element of vocabulary control). In an open-ended controlled vocabulary, the use of a preferred term as a substitute for one or more forbidden term of equivalent meaning is mandatory - it is binding on the searcher : other semantic links between terms are shown as optional routes for search expansion. Fixed vocabularies tend to have a higher percentage of mandatory devices than open-ended vocabularies, since some species/genus links are treated as mandatory.

9.3.3 It will here be assumed that all of the types of systems under review could be equipped with the full range of recall devices. Recall potential will, therefore, be eliminated as a factor in the present comparisons. Interest can, however, be focussed on a related issue ; how does the way a system is obliged, through the nature of its vocabulary, to adopt recall devices affect the effort required for indexing, searching and vocabulary maintenance?

9.4 Precision

9.4.1 Post-co-ordinate search situations

Where only post co-ordinate searching is possible, a system's precision performance is largely dependent upon the specificity of its terms; this situation favours open-ended controlled vocabularies and uncontrolled vocabularies. The latter have been shown to perform surprisingly well for SDI searches in the field of education[25]. In theory, their precision potential is lower than that of open-ended controlled vocabularies, if only because of their inability to distinguish between the different meanings of homographs. Of the various precision devices adopted by post-co-ordinate systems, roles and links have generally been shown to be ineffective, and expensive to apply[16]. Weighting, on the other hand - at the simple level of distinguishing between core terms and subsidiary terms - is almost certainly beneficial.

9.4.2 Searches of visually-readable indexes

In this area pre-co-ordinate systems have a clear advantage over post-co-ordinate ones, through their ability to show the searcher the various contexts in which his search terms have appeared. Two tests[12,17] have provided clear evidence to support the view that searchers can use the contextual information provided by pre-co-ordinate systems to avoid the retrieval of non-relevant documents and so achieve dramatic precision improvements over post-co-ordinate systems.

9.5 Search effort

9.5.1 Post-co-ordinate search situations

Uncontrolled vocabularies generally demand the greatest expenditure of effort in the compilation of search profiles, since the searcher must take account of all the variant ways in which the search concepts may be represented in the data base. In practice, the amount of effort required may be reduced if the system provides suitable search options e.g. a term truncation facility.

A procedure adopted by some post-co-ordinate hierarchically structured vocabularies to reduce the effort required to perform 'species' searches is 'upward posting'. By this procedure a search on a generic term retrieves all documents which are indexed by that term, or by any of its species. Upward posting is perhaps of most obvious value in feature card systems, where it may save the searcher much effort in card manipulation. It does not warrant further consideration in the present context, since, despite its use in ENDS[18] (the EURATOM Nuclear Documentation System), there are other more flexible ways of conducting 'species' searches in large machine-held files (i.e. the 'explode' facility in MEDLARS).

9.5.2 Searches of visually-readable indexes

Three points may be made about the types of index suited to visual searching :

a) Uncontrolled indexes of any kind (pre- or post-co-ordinate) are inappropriate - they require the searcher to carry out too many 'look-up'operations to compensate for the semantic scatter present in the index. Vickery has stated categorically 'Uncontrolled text words ... are suitable only for machine search'[19].

b) Post-co-ordinate indexes may also be said to be unsuited to this kind of search, not only because of their inferior precision potential to pre-co-ordinate indexes, but also on the grounds of search effort. They typically present the searcher with a list of postings under each term, and require him to perform Boolean operations on these lists (logical sum, product, and difference) to identify the document numbers satisfying his search prescription. These operations can be performed with great speed and accuracy by machine, but are highly time-consuming and error prone when carried out manually. In the opinion of this writer, the inappropriateness of using post-co-ordinate systems to provide indexes to printed bibliographies is clearly demonstrated in the index to the EUDISED R & D Bulletin[20].

c) Pre-co-ordinate controlled indexes are most suited to visual searching. Search time is reduced if :

i) the index consists of a single sequence, in which citations are entered directly under index entries (as in the BTI (British Technology Index) system)[21].

ii) entries are structured according to logical rules, so as to provide helpful subarrangement under each heading (as in PRECIS)

iii) the index provides entries under all of the significant term in each pre-co-ordinated string.

9.6 Habitability

9.6.1 In only one situation can a clear preference be made, on the grounds of habitability, for a particular type of indexing system. Where simple on-line searches are undertaken by uninitiated users, an uncontrolled system may prove the most habitable. Such a system allows the user to enter a search term without the need to consult a controlled vocabulary, and - as Lancaster has shown[22] - gives him a reasonable chance of retrieving at least a few relevant documents. There may well be sufficient to satisfy the user since, if only a portion of the database is available for on-line search (see 4.1 (a)) he will not in any case be expecting high recall.

9.6.2 In principle, there is no reason why a controlled system should not be as habitable as an uncontrolled one, provided that it is equipped with a full 'lead-in' vocabulary capable of mapping all forbidden terms entered by users into the corresponding preferred terms present in the date base.

9.7 Indexing effort

9.7.1 There is little doubt but that one of the major attractions of uncontrolled systems is that they can be applied with the minimum of indexing effort. In the simplest case - a 'free text' system - no indexing is required

at all, all significant words in titles and/or abstracts being accepted as search keys. Where human indexers are used either for 'free indexing' or for 'word-extraction' indexing they are able to work quickly, being free of the constraints of a controlled vocabulary or controlled syntax.

9.7.2 Where controlled systems are in use, post-co-ordinate indexing is generally faster and easier than pre-co-ordinate indexing, since in the latter case, the terms selected for a document must be arranged into strings according to a preferred citation order. Moreover, in post-co-ordinate indexing, the recognition of peripheral topics in a document normally lead to the assignment of a few additional terms, whereas in pre-co-ordinate indexing it often requires the formulation of several new strings. In practice it proves to be hardly feasible to use pre-co-ordinate systems for highly exhaustive indexing (say, an average of more than 15 terms per document).

9.7.3 Some pre-co-ordinate controlled systems possess a number of features which are designed to reduce indexing effort, e.g.

a) a 'string input' facility whereby the indexer is only required to formulate one string for each subject indexed, this string being manipulated by program to provide an index entry under each of the significant terms it contains

b) a mechanism, whereby the complete network of 'See' and 'See also' references appropriate to any term need only be recorded once and can therefore be called up as and when required, e.g. PRECIS' Reference Indicator Numbers (RINs).

c) a mechanism whereby, once the complete set of indexing data has been created for a particular document, that data may be automatically added to the records of any subsequent documents which deal with the same subject, e.g. PRECIS' Subject Indicator Numbers (SINs).

9.8 Vocabulary maintenance effort

9.8.1 The problems of vocabulary maintenance are in inverse proportion to the degree of vocabulary control. Fixed vocabularies present few problems, unless they are subject to constant changes as new editions are issued.

9.8.2 With open-ended controlled vocabularies, procedures are required.

a) for continuously updating the vocabulary, as new concepts are encountered in indexing

b) for notifying indexers and searchers of all new terms added.

These tasks prove particularly arduous when the vocabulary is 'young' and has a high growth rate, but become much less time-consuming once the core vocabulary has been established, and the number of new terms added begins to tail off.

9.8.3 If, as was earlier assumed, even an uncontrolled system needs a full range of recall devices, an efficient procedure is required for managing its vocabulary, so as to record all of the various terms by which a particular concept may be represented in the data base. The major problem with an uncontrolled vocabulary is essentially one of size - particularly if the terms it contains consist of phrases rather than individual words. INSPEC's experience in using 'free language' phrases may be cited[11] : 29,500 documents, indexed at an average exhaustivity of 6.5 phrases each, gave rise to a total vocabulary of more than 80,000 unique phrases.

9.9 Summary

The table below indicates the types of system considered 'best' in relation to each of the six performance criteria :

		SYSTEM WITH HIGHEST PERFORMANCE POTENTIAL
For high recall		All systems potentially equal
For high precision	Post-coordinate search situations	Open-ended controlled vocabularies, uncontrolled vocabularies
	Searches of visually-readable indexes	Pre-co-ordinate systems
For low search effort	Post-coordinate search situations	Controlled vocabularies
	Searches of visually-readable indexes	Pre-co-ordinate systems with controlled vocabularies
For habitability		Uncontrolled vocabularies (for certain types of on-line search, see 9.6.1)
For low indexing effort		Uncontrolled vocabularies
For low vocabulary maintenance effort		Fixed vocabularies

10. Browsing schemes

10.1 In Section 6.1 it was envisaged that current awareness bulletins and recurrent bibliographies would make provision for browsing searches. The following conditions are either desirable or necessary for browsing to be possible :

a) citations should be arranged in groups, in such a way that each group represents a subject or form of document of possible interest to users.

b) the groups should be arranged in a helpful order, which reflects the consensus view of the broad structure of the subject field of education (if such a view can be determined).

c) each citation should contain, or be accompanied by, an explicit statement of the subject content of the document to which it relates. The purpose of this statement is to help the browser decide whether or not the document is likely to be of interest to him.

a) and b) can be provided by a 'browsing scheme' i.e. a classification scheme or a system of subject headings. Titles might serve as the subject statements referred to in c), though abstracts would be preferable. Alternatively, each citation might be equipped with a specially constructed 'feature heading' derived from a pre-co-ordinated string of terms (cf the PRECIS-derived feature headings attached to citations in the 'British National Bibliography').

10.2 The view adopted here is that the functions of browsing schemes are as limited as indicated above. They are not required :

a) to provide for specific retrieval searches; this is the function of subject indexes

b) to provide each citation with a unique identifying number which can serve as the link between the citation and its subject index entries. Such a link is required but is preferably made independent of the browsing scheme. (It is appropriate to note, at this point, the practice adopted in the annual volumes of 'Library and Information Science Abstracts' : citations are arranged in classified order, but are also given a simple running number which serves as the link between index entries and citations).

10.3 A notional strategy will be suggested below for the use of a browsing scheme in the form of a simple enumerative classification. The scheme might follow the overall structure of an existing classification for the field of education e.g. the London Education Classification[23]. The plan envisaged here requires that each of the class numbers in the scheme be associated with several subject headings; each heading would be in a different language, but all would express the subject represented by the class number. A multi-lingual thesaurus containing a set of subject headings for each class number in the browsing scheme would be available in both printed and machine-readable form. Class numbers would be added to EUDISED records as part of their subject data. When a batch of records was processed to produce a bibliography or current awareness bulletin, the class number occurring in that batch would be input to a program which would access the multi-lingual thesaurus and extract the corresponding subject headings (in whatever language had been specified). Two courses of action are now possible :

a) create a classified file, in which each class number is accompanied by an explanatory subject heading

b) create an alphabetical subject sequence, in which class numbers are discarded and citations are arranged directly under subject headings

This strategy is of course highly tentative. It does, however, illustrate the possibilities of using a flexible browsing scheme capable of providing both alphabetical and classified arrangements, and sensitive to the need for multi-lingual access.

11. Multi-lingualism

11.1 It is obviously possible to provide multi-lingual access to documents by first abstracting and indexing them in one language, and then translating all abstracts and index terms into several other languages. However, this is an extremely uneconomical procedure. The only viable approach to the problems of multi-lingualism in a large scale system, such as EUDISED, lies in the development of trans-lingual procedures, i.e. procedures for 'switching' terms from one language to another, either automatically, or, by methods which require only minimal human intervention.

11.2 Translingualism may be attempted at a number of levels :

a) the translingual switching of subject headings in a browsing scheme

b) the translingual switching of terms in a post-co-ordinate system

c) the translingual switching of pre-co-ordinated strings of terms

d) the automatic translation of abstracts

e) the automatic translation of texts

a) has already been touched on, and will not be discussed further; e) is a topic which lies outside the scope of this paper. The leaves b) - d) for further consideration.

11.3 The translingual switching of post-co-ordinate terms can be achieved in two ways :

a) the direct equivalence approach, based upon a multi-lingual thesaurus, in which one-to-one equivalences are established between the terms of each language. This approach has the advantage of simplicity: the terms of one language are directly convertible to the terms of another. It brings with it the disadvantage that to achieve direct convertibility between terms, any specific terms in one language which do not have exact counterparts in all of the other languages are omitted from the thesaurus. This practice imposes an artificial limit on the specificity of terms, and, hence, reduces the precision potential of the thesaurus.

b) the 'switching language' approach, which entails the development of :

i) a 'switching language' containing a language-independent 'concept number' for each concept indexed (regardless of whether or not this concept can be expressed specifically in any particular language)

ii) two types of conversion table for each language

- a term-to-concept-number conversion table
- a concept-number-to-term conversion table

It is envisaged that in a decentralised multi-lingual network, conversion tables would be used as follows :

i) each centre in the network would index in its own language at the maximum level of specificity

ii) before contributing its records to the network, a centre would automatically convert all 'local language' terms to concept numbers, by means of a term-to-concept-number conversion table

iii) on receiving a batch of records from an external source, a centre would use a concept-number-to-term conversion table to convert the concept numbers on the in-coming records into terms in the 'local language'

The advantage of this approach is that it in no case interferes with indexing specificity.

11.4 The trans-lingual switching of pre-co-ordinate strings of terms presents special problems. Not only must it be possible to switch each of the terms in a string from one language to another, but this must be accomplished without distorting the meaning of the strings as a whole. Much work in this area has yet to be done. A fruitful approach to the problem seems to lie in the use of pre-co-ordinate system with a generalised language-independent syntax (such as the BTI system, or PRECIS), in conjunction with a switching language.

11.5 So far as is known only the TITUS system[24] can justly claim to be capable or automatically translating abstracts. The abstracts prepared for TITUS may be written in any of four languages, English, French, German and Spanish, but must be phrased in a stylised manner, according to the rules of a restricted syntax. The system uses a 'switching language' to convert all abstracts to a series of language-independent codes. The abstracts are stored in this form but may be processed by program to give output in any of the four languages previously noted. There are two questions about the performance of TITUS which remain, as yet, unanswered :

a) does the restricted syntax demanded by the system seriously reduce the quality of its abstracts?

b) is the system too costly to operate? - the time required to write an abstract for TITUS is known to be much greater than that required for the preparation of a conventional abstract.

12. A proposed strategy for the provision of subject data on EUDISED records

12.1 Thompson[3] has suggested that the bibliographic analysis of EUDISED materials might be carried out at several levels. The degree of analysis appropriate to any particular type of material would be determined by its 'importance' (as judged by whatever criteria might be established). Three levels of subject analysis are proposed below, ranging from Level 1, the most superficial, to Level 3, the most thorough.

12.2 Level 1

a) Controlled indexing

All documents would be indexed pre-co-ordinately using a controlled vocabulary. Typically, a document would be assigned one string of, perhaps, five or six terms. (Multi- topical documents would, naturally, receive more than one string). Strings would be used:

i) to produce visually-readable pre-co-ordinate indexes

ii) as the basis for machine-searching

It would not be possible to guarantee high recall in all cases, because of the relatively low exhaustivity of indexing employed. A possible subsidiary used for pre-co-ordinate strings would be in the provision of 'feature headings' to aid browsing (see 10.1).

The system used for pre-co-ordinate indexing should possess :

i) an open-ended thesaurus incorporating the full range of recall devices. The thesaurus would serve as the basis for the 'See' and 'See also' references provided in visually-readable indexes. It would also assist in the construction of profiles for machine searching.

ii) the ability to provide an index entry under each significant term in a string

iii) a generalised syntax which :

- ensures that entries are structured consistently so as to promote collocation in pre-co-ordinate indexes
- offers a basis for developing a multilingual facility within the system. Any such development would naturally draw heavily on the existing EUDISED multilingual thesaurus, but the vocabulary of this thesaurus would need to be extended, and, in some cases, terms would need to be modified to fit into the framework of a pre-co-ordinate system

iv) various 'labour-saving' features aimed at minimising indexing effort, e.g.

- a 'string input' facility (9.7.3 a))
- an efficient mechanism for calling up the network of 'See' and 'See also' references appropriate to any term (9.7.3 b))
- an efficient mechanism for handling 'recurrent' subjects (9.7.3 c))

b) Uncontrolled indexing

The natural language titles occurring in the citation in the databse would be available as additional search keys for machine searching.

c) Classification

All documents would be assigned one or more class numbers from a broad emmerative classification (which might be modelled on an existing classification for the field of education, such as The London Education Classification). If the suggestion made in 10.3 is accepted, the classification would form one component of an integrated classification/subject heading system, in which each

class number was equated with several semantically equivalent subject headings in various languages. The class numbers assigned to EUDISED records would then be capable of providing bibliographic tools with two types of browsing facility : a classified sequence of citations, and a sequence arranged alphabetically by subject heading. Class numbers might also be used in machine searching with or without other search terms. Their purpose would be to restrict a search to a particular class of record or to identify a particular subset of the database in preparation for the production of a special purpose bibliography or current awareness bulletin.

12.3 Level 2

At this level, the indexing data assigned at Level 1 would be 'enriched' by additional terms selected from the vocabulary of the pre-co-ordinate indexing system.

The 'enrichment' terms (say, 5 or 6 per document) would be chosen so as to express concepts which were not indexed at Level 1, because of the low exhaustivity of indexing practised at that level.

For the sake of economy in indexing effort, enrichment terms would not be pre-co-ordinated into strings, and would, therefore, play no part in the production of visually-readable indexes. Their sole purpose would be to increase the indexing exhaustivity, and so improve the recall of post-co-ordinate machine searches. The database would preserve the distinction between Level 1 terms and enrichment terms so that the former could, if necessary, be given a higher for the purposes of machine searching.

12.4 Level 3

Abstracts would be prepared for all documents processed at this level, so providing a basis for 'free text' searching.

REFERENCES

1. Encyclopaedia of Linguistics, information and control. Oxford, Pergamon Press, 1969.

2. Mills, J. Some current problems of classification for information retrieval. The Classification Society Bulletin 1 (4) 1968 : 18-27

3. Thompson, G.K. Abstracting services in education and the social sciences. In EUDISED technical studies 1971. Strasbourg, Documentation Centre for Education in Europe, 1971.

4. Journal citation data bases. Library Network/MEDLARS Technical Bulletin No.76 August 1975.

5. Lancaster, F.W. The cost-effectiveness analysis of information retrieval and dissemination systems. Journal of the American Society for Information Science 22 (1) January-February 1971 : 12-27.

6. Snyder, M.B. et al. Methodology for test and evaluation of document retrieval systems : a critical review and recommendations. McLean, Virginia, Human Sciences Research Inc., 1966. PB 169572.

7. Watt, W.C. Habitability. American Documentation 19 (3) July 1968 : 338-351.

8. Wellisch, H. Subject retrieval in the seventies - methods, problems, prospects. In Wellisch, H. ed. Subject retrieval in the seventies : proceedings on an international symposium held at Center of Adult Education, University of Maryland, College Park, May 14 to 15, 1971 : Westport, Connecticut, Greenwood Publishing Company, 1972.

9. Fong, E. A survey of selected document processing systems. Washington, United States National Bureau of Standards, 1971. Technical note 599.

10. Price, N.H. et al. On-line searching of Council of Europe conventions and agreements : a study in bi-lingual document retrieval. Information Storage and Retrieval 10 (3/4) March-April 1974 : 145-154

11. Field, B.J. Development of an integrated indexing and classification system. London, INSPEC, 1972. Report No.R72/10.

12. Keen, E.M. The Aberystwyth index languages text. Journal of Documentation 29 (1) March 1973 : 1-35.

13. Austin, D. The development of PRECIS : a theoretical and technical history. Journal of Documentation 30 (1) March 1974 : 47-102.

14. Excerpta Medica Foundation. Excerpta Medica automated storage and retrieval program of biomedical information. Amsterdam, Excerpta Medica Foundation, 1969.

15. Viet, J. EUDISED multilingual thesaurus for information processing in the field of education. First English ed. Paris, The Hague, Mouton, 1974.

16. Lancaster, F.W. On the need for role indicators in post-co-ordinate retrieval systems. American Documentation 19 (1) January 1968 : 42-46.

17. Keen, E.M. Computer-produced indexes - comparison and evaluation. (Handout prepared for an Aslib course on computer-produced indexes, October 1975).

18. Vernimb, C.O. Abstracts on microfiche for on-line retrieval. In. New developments in storage, retrieval and dissemination of aerospace information. Neuilly sur Seine, NATO Advisory Group for Aerospace Research and Development, 1973.

19. Vickery, B.C. Structure and function in retrieval languages. Journal of Documentation 27 (2) June 1971 : 69-82.

20. EUDISED R & D Bulletin : experimental issue. Strasbourg, Documentation Centre for Education in Europe, 1975.

21. Coates, E.J. Computer handling of social science terms and their relationshi In EUDISED technical studies. vol.III. Strasbourg, Documentation Centre for Education in Europe.

22. Lancaster, F.W. et al. Evaluating the effectiveness of an on-line natural language retrieval system. Information Storage and Retrieval 8 (5) October 1972 : 223-245.

23. Foskett, D.J. The London Education Classification : a thesaurus/classificati of British educational terms. London, Institute of Education Library, University of London, 2nd ed. 1974.

24. Ducrot, J.M. Le Systeme TITUS II. Information et Documentation (4) October 1973 : 3-40.

25. Tell, B.V. et al. The use of ERIC tapes in Scandinavia, searching with thesaurus terms in natural language. Strasbourg, Documentation Centre for Education in Europe, 1972. ED 072794.